W454–W458 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 8 June 2013*
doi:10.1093/nar/gkt472

HIV N-linked glycosylation site analyzer and its further usage in anchored alignment

Timothy I. Shaw[1,2] and Ming Zhang[1,2,*]

[1]Institute of Bioinformatics, University of Georgia, GA 30605, USA and [2]Department of Epidemiology and Biostatistics, University of Georgia, GA 30605, USA

Received March 7, 2013; Revised May 1, 2013; Accepted May 7, 2013

ABSTRACT

N-linked glycosylation is a posttranslational modification that has significantly contributed to the rapid evolution of HIV-1. In particular, enrichment of N-linked glycosylation sites can be found within Envelope variable loops, regions that play an essential role in HIV pathogenesis and immunogenicity. The web server described here, the HIV N-linked Glycosylation Site Analyzer, was developed to facilitate study of HIV diversity by tracking gp120 N-linked glycosylation sites. This server provides an automated platform for mapping and comparing variable loop N-linked glycosylation sites across populations of HIV-1 sequences. Furthermore, this server allows for refinement of HIV-1 sequence alignment by using N-linked glycosylation sites in variable loops as alignment anchors. Availability of this web server solves one of the difficult problems in HIV gp120 alignment and analysis imposed by the extraordinary HIV-1 diversity. The HIV N-linked Glycosylation Site Analyzer web server is available at http://hivtools.publichealth.uga.edu/N-Glyco/.

INTRODUCTION

Strategic placement and loss and gain of N-linked glycosylation sites are one of the most important evolutionary mechanisms adopted by HIV-1 to generate its extraordinary sequence diversity (1). A typical N-linked glycosylation site requires the context of the amino acid pattern N-X-[S or T] (2), with X being any amino acid except Proline (3). Highly glycosylated regions are referred to as immunologically silent faces (4), reducing antigenicity and restricting access to chemokine receptors. Changes in N-linked glycosylation sites in HIV-1 can induce conformational changes in Envelope gp120, diminishing binding of many gp120-specific antibodies (5). Comparison between neutralization-sensitive and neutralization-resistant HIV-1 strains shows a higher number of glycosylation sites associated with the resistant clusters (6). Changes in N-linked glycosylation sites have also been linked to both disease stage and co-receptor usage. Leal *et al*. reported an increase in N-linked glycosylated sites during late stages of HIV-1 infection (7). Evaluation of co-receptor usage has demonstrated a tendency for higher mutation rates, higher net positive charges and fewer glycosylation sites within HIV-1 strains with CXCR4 co-receptor usage (8).

In HIV-1, N-linked glycosylation sites are enriched within the variable loops, which contain multiple neutralizing antibody-binding sites (9). Changes of N-linked glycosylation sites within variable loops, as well as changes of lengths of variable loops imposed by frequent indels (insertion and deletions), are highly favored in HIV-1 (1,9). Both changes are important measurements of HIV-1 diversity (10). Of note, although immunologically and evolutionarily important, HIV-1 variable loops are notoriously known as difficult to analyze owing to extraordinary viral diversity in these regions (11). As a result, variable loops are typically excluded from phylogenetic analyses (6,7,10), leading to frequent underestimation of HIV-1 diversity in immunologically important genomic regions.

To address the importance of N-linked glycosylation sites in HIV-1 and problems in analyzing variable loops as described above, we present development of the HIV N-linked Glycosylation Site Analyzer, available at http://hivtools.publichealth.uga.edu/N-Glyco/. This server provides an automated platform for mapping and comparing N-linked glycosylation sites within variable loops between populations of HIV-1 sequences. Furthermore, considering the functional importance and conserved patterns of N-linked glycosylation sites, we have implemented in this server a feature that optimizes HIV-1 sequence alignment using N-linked glycosylation sites in variable loops as alignment anchors. As a result, our N-linked Glycosylation Site Analyzer serves as a valuable gateway for exploring HIV-1 diversity in immunologically important genomic regions, contributing to an improved understanding of host–virus interaction and enhanced viral vaccine strain selection.

*To whom correspondence should be addressed. Tel: +1 706 542 2194; Fax: +1 706 583 0695; Email: mzhang01@uga.edu

MATERIALS AND METHODS

Two key features distinguish our HIV-1 N-linked Glycosylation Site Analyzer from other HIV-1 sequence analysis tools and servers. First, through an automated pipeline, changes at N-linked glycosylation sites within each variable loop region, as well as loop lengths, can be easily tracked and compared between populations of HIV sequences. Second, the server optimizes HIV-1 sequence alignment by using the N-linked glycosylation site as alignment anchor. Implementation of both features has been written in Java. The web server interface is implemented through HTML and Bootstrap JavaScript. Visualization methods are available for all results (see details in 'Server Output' section below).

Algorithm

In the N-linked Glycosylation Site Comparison program (N-Glyco Site Compare), input sequences are automatically aligned with the HIV-1 reference strain HXB2 (accession number: K03455. http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html). Through implementation of the HIV alignment algorithm as described by Gaschen *et al* (12), the variable loops V1–V5 are identified and clipped based on genomic coordinates defined in HIV Sequence Compendium 2012 (13). Within each variable loop region, the N-linked glycosylation sites, whose pattern is N-X-[S or T] (2), are identified by pattern matching of asparagine followed by any amino acid except Proline, followed by either a serine or threonine. In the case of continuous N-linked glycosylation sites (e.g. NNST), only the first N-linked glycosylation site is counted because two continuous N-linked glycosylation sites would induce steric occlusion. An exception exists for NNST in which the second glycosylated asparagine is counted because N-X-T are more frequently glycosylated than N-X-S (14), and oligosaccharyltransferase has a higher affinity for N-X-T than N-X-S (15).

In the program 'V Loop Alignment', we optimize V loop region alignments by using N-linked glycosylation sites within V loops as alignment anchors. The 'V Loop Alignment' program accepts input for both aligned and unaligned sequences. The HXB2 sequence (accession number: K03455) is used as the reference in the alignment procedure; therefore, HXB2 is automatically added to the input sequences when absent from the user input. For unaligned sequences, they are initially aligned through a HMMER-generated HIV profile (12). The aligned sequences, from direct user input or HMMER-derived alignment, will then be refined based on a heuristic approach for manual curation of HIV-1 alignments (6,16,17). For each variable loop region, the input sequence with the highest number of N-linked glycosylation sites for that region is identified, and its N-linked glycosylation sites used as alignment anchors for all input sequences. This process continues through each variable loop regions. The N-linked glycosylation sites for the rest of the input sequences are then aligned to these anchors based on a greedy algorithm, mapping each N-linked glycosylation site to its closest available anchor.

RESULTS

N-linked Glycosylation Site Comparison program

Input

The N-Glyco Site Compare program is designed to compare groups of HIV sequences for variation and changes in N-linked glycosylation patterns. The comparison groups are those sequences under different conditions, for instance, sequences at different time points, of different subtypes and associated with different risk factors. The N-Glyco Site Compare program reads in two sets of FASTA sequences, namely query and background, respectively, and compares their N-linked glycosylation site frequency and variable loop lengths. Three options are provided for selecting the background sequences: (i) No background, which allows N-glycosylation site analysis to be performed in one single sequence or one set of sequences (i.e. in the query set); (ii) Using the most recent HIV-1 M group reference sequence set as the background. The reference sequences were obtained from the Los Alamos HIV Sequence Database group; and (iii) User-defined background sequences, which bestow flexibility in performing user-defined comparisons. The input of the N-Glyco Site Compare program can be either aligned or unaligned gp120 sequences. Both nucleotide and protein sequences are acceptable as input.

Output

Output from the N-Glyco Site Compare program highlights N-linked glycosylation sites through a graphical histogram spanning across HXB2 Envelope positioning (18) (Figure 1A). Loop length and frequency of N-glycosylation site distribution within each variable loop (V1–V5) are compared and depicted in a boxplot between comparison groups (Figure 1B and C). Furthermore, a two-sided Wilcoxon test with 1000 times of Monte Carlo resampling is provided for comparison statistics. The N-Glycosylation site and V loop mapping for each sequence are provided. Visual representation for the N-Glycosylation mapping is described in further detail in the section below ('N-linked Glycosylation Site Alignment Program—Output' section).

N-linked Glycosylation Site Alignment program

Input

The N-linked Glycosylation Site Alignment program ('V Loop Alignment' program) reads in one FASTA input file, regardless aligned or not. Both nucleotide and protein sequences are acceptable. The N-glycosylation sites within the variable loops are used as alignment anchors to optimize sequence alignments as described in the Algorithm Section.

Output

Two mechanisms are used for visualizing the N-linked glycosylation optimized alignment: (i) Jalview, a Java-based alignment editor that provides extensive

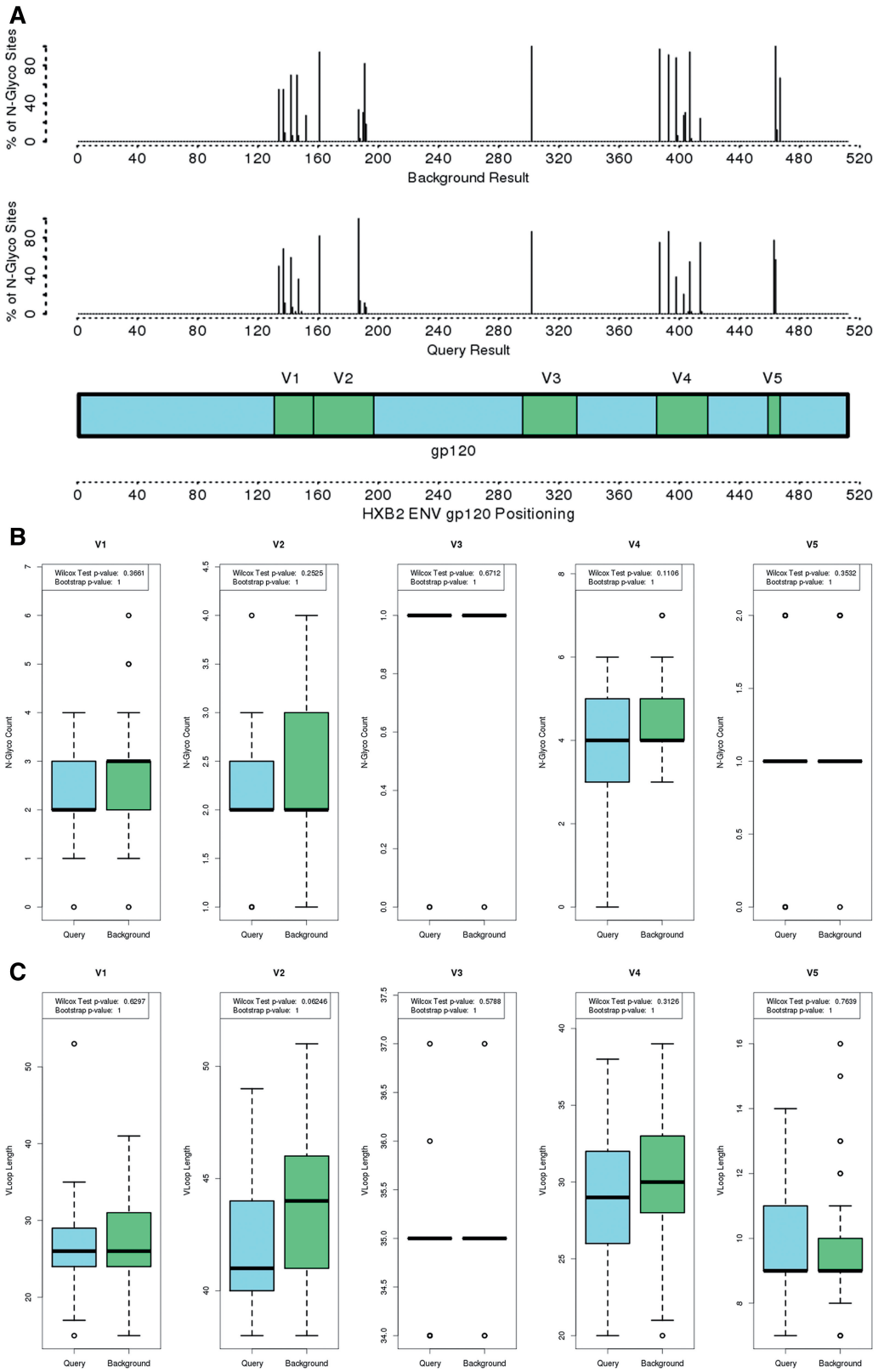

Figure 1. An example output of N-Glyco Site Compare program. (**A**) Location of identified N-linked glycosylation sites within the variable loops (V1–V5) in terms of HXB2 numbering (http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html). Y-axis: Percentage of sequences with N-linked glycosylation site at each alignment position. X-axis: HXB2-based gp120 sequence positions. (**B** and **C**) The distribution of number of N-linked glycosylation sites and lengths of variable loops. *P*-value is calculated in two-sided Wilcoxon test. The bootstrap *P*-value is calculated by 1000 times of Monte Carlo resampling.

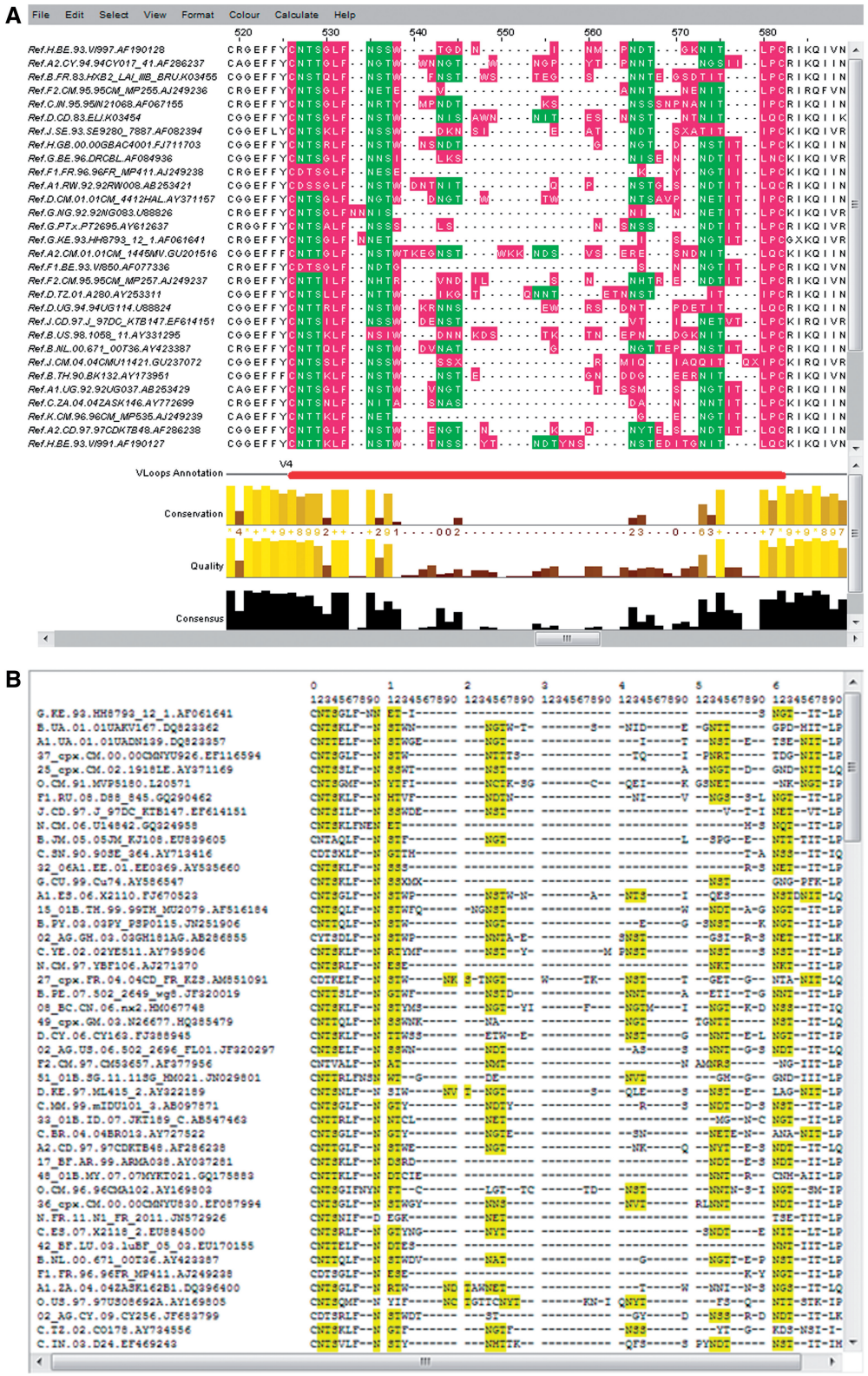

Figure 2. An example output of N-linked Glycosylation Site Alignment program. (**A**) A Jalview-based alignment editor depicts an optimized alignment using N-linked glycosylation sites as alignment anchors. From the V loop Alignment program, V loop regions are highlighted as pink and each N-linked glycosylation sites are highlighted in green. An additional V loop annotation track is added underneath the alignment. (**B**) An HTML view of the optimized alignment with the N-linked glycosylation sites are highlighted in yellow.

functionality for alignment visualization and editing (19,20). The V loop regions are highlighted as pink and N-linked glycosylation sites are highlighted in green; an additional V loop annotation track is added underneath the alignment.(Figure 2A); and (ii) an HTML-based visualization of the alignment with N-linked glycosylation sites highlighted in yellow (Figure 2B). The nucleotide and protein version of the alignment are downloadable. Also available in the downloadable results are the annotation for the location of N-linked glycosylation site and V loop region for each sequence.

CONCLUSION

Our HIV-1 N-linked Glycosylation Site Analyzer provides an automated platform to map and compare patterns of N-linked glycosylation sites between populations of HIV-1 sequences. In addition, to address the problem of improper variable loop region alignment that causes underestimation of HIV-1 diversity, we have developed an algorithm for performing anchored alignment based on N-linked glycosylation sites. The toolset and analysis pipeline described here can be extended to understanding diversity and N-linked glycosylation patterns in other viruses. Our web server provides an important gateway to track N-linked glycosylation site patterns within HIV-1 populations, thus improving our capability to better understand viral diversity under changing contexts of antigenic structures and transmission mechanisms.

ACKNOWLEDGEMENTS

The authors thank Ms Tess Z. Griffin and anonymous reviewers for help and comments on the manuscript. We would also like to thank members of the Ming Zhang HIV Lab and various HIV research groups for their extensive testing of our web server.

FUNDING

NIH [R03AI104258]; University of Georgia Research Fund [10793GR002] and University of Georgia Research Foundation Award [1021RX064536]. We would also like to acknowledge the support from the ARCS foundation for TIS. Funding for open access charge: University of Georgia [UGA10793GR002].

Conflict of interest statement. None declared.

REFERENCES

1. Zhang,M., Gaschen,B., Blay,W., Foley,B., Haigwood,N., Kuiken,C. and Korber,B. (2004) Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*, **14**, 1229–1246.
2. Marshall,R.D. (1974) The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. *Biochem. Soc. Symp.*, 17–26.
3. Gavel,Y. and von Heijne,G. (1990) Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng.*, **3**, 433–442.
4. Moore,J.P. and Sodroski,J. (1996) Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein. *J. Virol.*, **70**, 1863–1872.
5. Si,Z., Cayabyab,M. and Sodroski,J. (2001) Envelope glycoprotein determinants of neutralization resistance in a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2) derived by passage in monkeys. *J. Virol.*, **75**, 4208–4218.
6. Kulkarni,S.S., Lapedes,A., Tang,H., Gnanakaran,S., Daniels,M.G., Zhang,M., Bhattacharya,T., Li,M., Polonis,V.R., McCutchan,F.E. *et al*. (2009) Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C HIV-1 Env clones from India. *Virology*, **385**, 505–520.
7. Leal,E., Casseb,J., Hendry,M., Busch,M.P. and Diaz,R.S. (2012) Relaxation of adaptive evolution during the HIV-1 infection owing to reduction of CD4+ T cell counts. *PLoS One*, **7**, e39776.
8. Lin,N.H., Becerril,C., Giguel,F., Novitsky,V., Moyo,S., Makhema,J., Essex,M., Lockman,S., Kuritzkes,D.R. and Sagar,M. (2012) Env sequence determinants in CXCR4-using human immunodeficiency virus type-1 subtype C. *Virology*, **433**, 296–307.
9. Wyatt,R., Kwong,P.D., Desjardins,E., Sweet,R.W., Robinson,J., Hendrickson,W.A. and Sodroski,J.G. (1998) The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*, **393**, 705–711.
10. Korber,B., Gaschen,B., Yusim,K., Thakallapally,R., Kesmir,C. and Detours,V. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.*, **58**, 19–42.
11. Abecasis,A., Vandamme,A.M. and Lemey,P. (2007) Sequence alignment in HIV computational analysis. In: Thomas Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B (eds). *HIV Sequence Compendium 2006|2007*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 07-4826, pp. 2–16.
12. Gaschen,B., Kuiken,C., Korber,B. and Foley,B. (2001) Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics*, **17**, 415–418.
13. Kuiken,C., Leitner,T., Hahn,B.H., Mullins,J.I., Wolinsky,S., Foley,B., Apetrei,C., Mizrachi,I., Rambaut,A. and Korber,B. (2012), *HIV Sequence Compendium 2012*. T-6 Group, MS K710, Los Alamos National Laboratory, NM, USA. LA-UR-12-24653.
14. Kasturi,L., Eshleman,J.R., Wunner,W.H. and Shakin-Eshleman,S.H. (1995) The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. *J. Biol. Chem.*, **270**, 14756–14761.
15. Gerber,S., Lizak,C., Michaud,G., Bucher,M., Darbre,T., Aebi,M., Reymond,J.L. and Locher,K.P. (2013) Mechanism of bacterial oligosaccharyltransferase: in vitro quantification of sequon binding and catalysis. *J. Biol. Chem.*, **288**, 8849–8861.
16. Gnanakaran,S., Bhattacharya,T., Daniels,M., Keele,B.F., Hraber,P.T., Lapedes,A.S., Shen,T., Gaschen,B., Krishnamoorthy,M., Li,H. *et al*. (2011) Recurrent signature patterns in HIV-1 B clade envelope glycoproteins associated with either early or chronic infections. *PLoS Pathog.*, **7**, e1002209.
17. Zhang,M., Foley,B., Schultz,A.K., Macke,J.P., Bulla,I., Stanke,M., Morgenstern,B., Korber,B. and Leitner,T. (2010) The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology*, **7**, 25.
18. Korber,B., Foley,B.T., Kuiken,C., Pillai,S.K., and Sodroski,J.G. (1998) Numbering positions in HIV relative to HXB2CG. In: Korber,C.K., Foley,B., Hahn,B., McCutchan,F., Mellors,J. and Sodroski,J (eds). *Human Retroviruses and AIDS 1998*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, pp. III-102-111.
19. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
20. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The jalview java alignment editor. *Bioinformatics*, **20**, 426–427.